Linear Representations of Sentiment in Large Language Models

Year: 2023 | Citations: 119 | Authors: Curt Tigges, O. Hollinsworth, Atticus Geiger

---

Abstract

Sentiment is a pervasive feature in natural language text, yet it is an open question how sentiment is represented within Large Language Models (LLMs). In this study, we reveal that across a range of models, sentiment is represented linearly: a single direction in activation space mostly captures the feature across a range of tasks with one extreme for positive and the other for negative. Through causal interventions, we isolate this direction and show it is causally relevant in both toy tasks and real world datasets such as Stanford Sentiment Treebank. Through this case study we model a thorough investigation of what a single direction means on a broad data distribution. We further uncover the mechanisms that involve this direction, highlighting the roles of a small subset of attention heads and neurons. Finally, we discover a phenomenon which we term the summarization motif: sentiment is not solely represented on emotionally charged words, but is additionally summarized at intermediate positions without inherent sentiment, such as punctuation and names. We show that in Stanford Sentiment Treebank zero-shot classification, 76% of above-chance classification accuracy is lost when ablating the sentiment direction, nearly half of which (36%) is due to ablating the summarized sentiment direction exclusively at comma positions.